模倣レポート判定に用いる文書間類似度の考案

太田 貴久　　増山 繁
豊橋技術科学大学 知識情報工学系
{kikyu, masuyama}@smlab.tutkie.tut.ac.jp

1　はじめに

今日，コンピュータ，インターネットの普及により，電子化された文書の取得，編集は非常に容易となっている．これにより，他人のレポートをコピー&ペーストして，それに多少の手直しを加えて提出する (以下，このように生成されたレポートを「模倣レポート」と呼ぶ) ことが問題となっている．そこで，レポートの模倣の程度を定量化し，それを判別するシステムを作ることが出来れば，自動的に模倣レポートを判別できるだけでなく，その存在自体が模倣レポートの抑制に繋がると考えられる．

これまでに模倣レポートの判別を目的とした研究は少なく，文献 [1], [2], [3] くらいであまりない．文献 [1] では n-gram 解析を用いた類似度計算法を提案し，文献 [2] では $tf \cdot idf$ を用いたベクトル空間法による類似度計算法を提案している．これらの手法では文書全体を単語の集合として取り扱っている．そのため，同一のテーマについて書かれている文書を模倣ではないにも拘わらず模倣であると誤識別する可能性があり，さらに，部分的に模倣している文書では正しく判別されないと考えられる．また，文献 [3] では文書を文単位に分割しているが，文の結合／分割が発生した場合に本質的に対処ができないという問題点がある．

本手法では構文情報を用いて，文間類似度を求め，そこから文書間類似度を求める．文書間類似度の計算に構文情報を用いる研究には文献 [5],[6] などがあるが，これらの手法は一般的な文書間の類似度を求める手法であり，模倣した文書を対象とすることを前提とした本研究とは本質的に異なる．

以下では，まず手法考案に当たって模倣レポートを調査し，そこに使われている技法を分類した．そして，それらの技法に対して有効に働く文書間類似度の計算法を提案し，最後に実験用のレポート集合にその手法を適用し，その結果と考察を述べる．

2　手法考案にあたって

今回，実験のため 5 名 (模倣レポート作成経験者) に模倣レポートを作成してもらった．それらを分析することにより，模倣レポートに使われる技法を 3 つのレベルに分類した．それを表 1 に示す．

ここで，表層レベルの語調の変換とは，です／ます調から，だ／である調への変換などを表す．また，構文レベルの文節の削除／追加とは冗長な修飾節を加えたり削除したりすることである．このうち，模倣レポートに使われる技法はレベル 1 とレベル 2 が多く，レベル 3 の意味レベルでの模倣は殆ど見受けられなかった．これより，本研究で取り扱うのはレベル 1~2 の技法とした．

また，レポートにはテーマが与えられているので，必然的に同一の分野について書かれていることが多い．これら同一分野の文書を模倣と判定することは極力回避しなければならない．

表 1: 模倣のレベル

レベル	使われる技法
表層レベル (レベル 1)	語調の変換 記号の変換 (句点からピリオドに等) 換言 (動詞から名詞＋する等)
構文レベル (レベル 2)	文節の削除／追加 文節の入れ替え 文の分割／結合
意味レベル (レベル 3)	文章の再構成 (段落単位の入れ替え等) 文の追加 (独自な意見の追加等)

さらに，現実にレポート判別システムを構築し，利用することを考えたとき，最終的な評価は人間が行うので，誤識別は出来るだけ少ない方が望ましく，再現率より精度を優先するべきである．

これより，本研究で目指すべき模倣レポート判別用の文書間類似度は以下のような特徴を持つことが望ましい．

- 単に同一分野について書かれている文書を誤識別しない
- 表層レベル，構文レベルの違いは模倣と判定する
- 再現率より精度を優先する

そこで，これらを満たすような手法を考案する．

3　文書間類似度計算法

文書の類似度を求めるにあたり問題となるのは，「どのような特徴」を「どのように計算するか」である．文書は文を繋げたものである．また，文は単語 (形態素) を繋げたものである．これより，文書の「意味」は単語が繋がることで発生すると考えた．そこで，本手法では文書を文に分割し，さらに文を独自に定義した最小構成単位に分割し，その類似度から文書の類似度を計算する．そして，文の類似度を計算する際，著者の記述に関する特徴，すなわち文体を除去し，文の内容のみを評価する必要がある．これは，助詞，助動詞が文体を表すとの従来の研究 [4] から，まず，それらを除去し，さらに，副詞や連体詞などの付属語も文の内容には直接関係しないと判断して除去した．最終的に，自立語である動詞，形容詞，名詞のみを処理の対象とすることで，文体の要素を完全に取り除くことが出来ると考えた．

以下で文書間類似度の計算手順を述べる．

3.1　文書間類似度

文書 d_x と d_y の類似度 $\widehat{sim}_d(d_x, d_y)$ を次のように定義する．

$$\widehat{sim}_d(d_x, d_y) = \frac{2sim_d(d_x, d_y)}{sim_d(d_x, d_x) + sim_d(d_y, d_y)}$$

$$sim_d(d_x, d_y) = \sum_{s_m \in d_x} \sum_{s_n \in d_y} sim_s(s_m, s_n)$$

ここで, s_m : 文書 d_xの m 番目の文
s_n : 文書 d_yの n 番目の文
である.

各文の類似度の総和を文書の自己類似度で正規化したものを文書間類似度とした．このように，文間類似度から文書間類似度を求めることで，部分的に模倣しているレポートも発見することが出来る.

3.2 文間類似度

文間類似度 $sim_s(s_m, s_n)$ を次のように定義する.

$$sim_s(s_m, s_n) = \sum_{ms_i \in s_m} \sum_{ms_j \in s_n} \widehat{sim}_{ms}(ms_i, ms_j)$$

ここで，ms は独自に定義した「最小構成文」であり，これは文から意味を成す最低限の語を抜き出したものである．例を 3.3.1 に示す．そして，文間類似度を最小構成文間類似度の総和とする.

最小構成文の生成と，その類似度計算について以下で説明する.

3.3 最小構成文間類似度

3.3.1 最小構成文の生成

最小構成文は次のように生成する.

Step 1 構文解析を行い，係り受け木を生成する.
Step 2 Step 1 で生成した木の葉から根への全てのパスを抽出し，形態素列へ変換する.
Step 3 各パスから名詞，動詞，形容詞以外の形態素を削除する.

具体的な例を挙げて説明する．例として以下の文の最小構成文生成を考える.

例文：EC が 12 月 31 日で市場の統合を完成させる.

この文の係り受け木は図 1 のようになる.

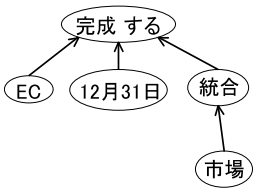

図 1: 例文の係り受けグラフ

そして，上記の方法で抽出される最小構成文は次の 3 つになる.

ms_1 : EC, 完成, する
ms_2 : 12 月 31 日, 完成, する
ms_3 : 市場, 統合, 完成, する

このような方法で文を分割すると係り受け木の根に近い形態素ほど多く評価されることになる．これは，根に近い語ほど意味の上で重要な語であると判断したからである.

3.3.2 最小構成文間類似度の計算

最小構成文間類似度を以下の式で定義する.

$$\widehat{sim}_{ms}(ms_i, ms_j) = \frac{sim_{ms}(ms_i, ms_j) + sim_{ms}(ms_j, ms_i)}{2}$$

$$sim_{ms}(ms_i, ms_j) = \sum_{k=1}^{|ms_i|} rel(C_{ms_j}(w_{ms_i,k-1})), C_{ms_j}(w_{ms_i,k}))$$

$$rel(w_1, w_2) = \begin{cases} \lambda^{d(w_1,w_2)}, & w_1, w_2 \text{ が存在するとき} \\ 0, & \text{それ以外のとき} \end{cases}$$

(ただし, λ は $0 < \lambda \leq 1$ を満たす任意の定数)

$w_{ms,k}$: 最小構成文 ms の k 番目の語
$C_{ms}(w)$: 最小構成文 ms 中で語 w に対応する語
$d(w_1, w_2)$: 語 w_1 と語 w_2 の間に存在する語の数

本手法で用いる語 (形態素) の意味的な類似度は一致するか，しないかの 2 値である．さらに，最小構成文上で 2 語以上一致する語が存在しなければ文の類似度が発生しないように定義している．このようにすることで，模倣でない文書の類似度を低く抑えることができ，これは精度の向上につながる．今回は形態素の基本形と品詞が完全に一致したとき，語が対応すると判定した．また，sim_{ms} の計算は非対称なので，文間類似度には対称化した最小構成文間類似度 $\widehat{sim}_{ms}$ を用いる.

例として，$\lambda = 0.5$ として，以下の 2 つの対称化前の最小構成文の類似度 $sim_{ms}(ms_1, ms_2)$ の求め方をあげる.

ms_1 : EC, 完成, する
(上記例文より)
ms_2 : EC, 市場, 統合, 完成, する
(原文：「EC の市場統合が完成する」)

これに上記の式を当てはめると次のようになる.

$$\begin{aligned} sim_{ms}(ms_1, ms_2) &= rel(C_{ms_2}(\text{EC}), C_{ms_2}(\text{完成})) \\ &\quad + rel(C_{ms_2}(\text{完成}), C_{ms_2}(\text{する})) \\ &= 0.5^2 + 1 = 1.25 \end{aligned}$$

最初に ms_1 の 1 番目の語「EC」と 2 番目の語「完成」に着目する．これらの語を ms_2 から探し出す．すると,「EC」は 1 番目,「完成」は 4 番目にある．このとき,「EC」と「完成」の間には 2 語存在するので $0.5^2 = 0.25$ となる．同様に ms_1 の 2 番目と 3 番目の語「完成」と「する」を ms_2 から探し出すと，今度は 4,5 番目に発見することができ $0.5^0 = 1$ となる．最後にこの 2 つを足し合わせた値 1.25 が最小構成文間の類似度となる.

3.3.3 並列・同格の関係について

並列・同格が発生した場合，並列・同格が発生した点で木を内包するノードが生成され，内部ノードの数だけパスが分岐するものとする.

例として以下の文を考える.

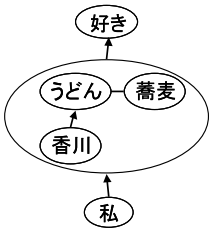

図 2: 並列・同格の例

例文 ： 私は，蕎麦と香川のうどんが好きだ．

この文の係り受け木は図 2 のようになる．
そして，この木から生成される最小構成文の集合は以下のようになる．

ms_1 ： 私, 蕎麦, 好き
ms_2 ： 私, 香川, うどん, 好き

文書間の類似度に構文情報を用いるものとしては文献 [5],[6] がある．これらの手法との違いは，文献 [5] とは木をパスに分解して語の入れ替えに対応している点が異なり，文献 [6] とは語の類似度を用いていない点で異なる．

4 実験

今回，実験のため,「テレビゲームの中古販売」について書かれたレポートを 3 つ (文書 1~3) と，文書 1 を被験者に「人のレポートを写すつもりで，読みながら写せ」と指示して作成した模倣レポートを 2 つ (文書 4,5) の計 5 文書を用意した．そして，以下の 2 つの実験を行った．

- **実験 1**
  実験用の文書集合に対して λ を 0.1 から 1.0 まで変化させて，その類似度の変化を調べた．
- **実験 2**
  文書 1 と文書 4 の組と，文書 1 と文書 3 の組についてそれぞれ文間類似度の分布を調べた．

今回，形態素解析器には JUMAN[1]，構文解析器には KNP[2]を使用した．

5 結果

5.1 実験 1

文書 1 とそれ以外の文書 (文書 2~5) との類似度の変化を図 3 に示す．このように，λ を変化させてもあまり結果は変わらない．また，$\lambda = 0.5$ の各文書間類似度を表 2 に示す．ここで，模倣レポートと，そのオリジナルの組である文書 1-4 の類似度はかなり高いものとなっており，明らかに模倣レポートであると判別できる．しかし，文書 1-5 の組の類似度は模倣でないレポートよりは高いが，はっきり判別できない値となってしまっている．これは文書 5 には換言が多く使われていたためである．レポートの模倣においてよく使われる換言は「動詞」から「名詞＋する」への変換であった．この変換が行われる場所は係り受け木の根であることが多い．よって，係り受け木の根の方を多く評価し，換言に未対応の本手法では換言が多く使われている文書では低い類似度となってしまう．

[1]http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html
[2]http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html

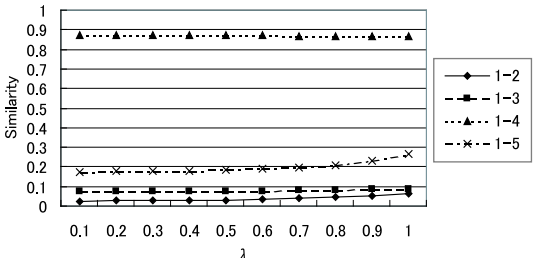

図 3: λ の変化による類似度の推移

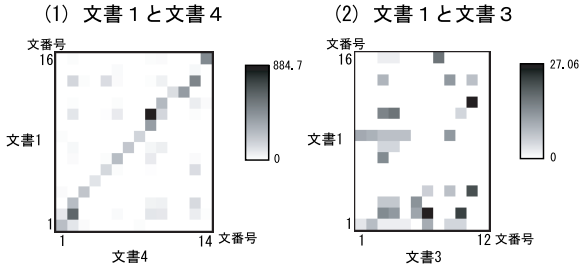

図 4: 文間類似度の分布

表 2: $\lambda = 0.5$ のときの結果

文書対	類似度	文書対	類似度	文書対	類似度
1-2	0.029	2-3	0.033	3-5	0.070
1-3	0.084	2-4	0.030	4-5	0.203
1-4	0.865	2-5	0.008		
1-5	0.192	3-4	0.100		

5.2 実験 2

文間類似度の分布を図 4 に示す．本手法では文間類似度の正規化は行っていないことから，文間類似度の分散は非常に大きくなる可能性がある．そこで，今回は分布の特徴を掴むために文間類似度を $\sqrt{1 - \frac{(sim/max(sim))^2}{4}}$ (楕円の方程式) で正規化を行っている．ここで，sim は対象としている文間類似度で，$max(sim)$ は文間類似度の最大値である．
図 4 より，模倣レポートとの文間類似度の分布には明らかな相関が見られ，逆に，模倣でないレポートとの文間類似度の分布には相関は見られないことが分かる．

6 考察

結果より，模倣したレポートと，そのオリジナルの間では高い類似度を得ることができた．しかし，換言が多く使われたレポートと模倣ではない文書間の類似度の差があまりない．これを解決する方法としては文書間類似度の計算方法の変更がある．本手法では文書間類似度を文間類似度の総和で求めている．図 4-(1) からも明らかなように，模倣レポートの文

間類似度の分布は 1 次関数に帰着できる．逆に，模倣していない (主題の異なる) レポートでは分布に規則性は見られない．これより，文書間類似度を，コストとして文間類似度を用いた動的計画法で求めた方が，より良い結果になると考えられる．

そして，文書 5 のような換言が多く使われている文書に対しての文書間類似度が低くなる問題に対しては，語の対応にシソーラスを用いることや，文献 [3] にあるような概念辞書を用いる方法で対処できると考えられる．

また，本手法は構文解析器の精度に大きく作用される．学生レポートのような日本語としての完成度が低い文書に対しては特にこの影響が大きいと思われるが，現在の構文解析器の精度でも十分な結果が得られている．

本手法や文献 [1], [2], [3] の手法のように模倣レポートの判別に文書間類似度という 1 次元の値を用いた場合,「全体がなんとなく似ている」文書と「部分的に似ている」文書を区別できない．実用化を考えるのならば,「どの部分」が「どれだけ似ているか」が分かる指標が必要である．

7 追加実験

考察でも述べたように，模倣レポートとの文書間類似度には明らかな相関が見られ，逆に，模倣でないレポートには相関は見られない．これより，文書間類似度に文間類似度をコストとした動的計画法を導入することで更に精度が上がるのではないかと考えた．そこで，文書間類似度に動的計画法を導入したものに対して実験 1($\lambda = 0.5$ の場合) を実験を行った．

7.1 結果

各文書対に対する文書間類似度の結果を表 3 に示す．

表 3: 動的計画法を用いた文書間類似度

文書対	類似度	文書対	類似度	文書対	類似度
1-2	0.017	2-3	0.025	3-5	0.047
1-3	0.060	2-4	0.020	4-5	0.151
1-4	0.872	2-5	0.005		
1-5	0.147	3-4	0.073		

7.2 考察

実験用のレポート集合では文書間類似度に動的計画法を導入することで，わずかではあるが模倣でないレポートの文書間類似度を抑制できている．しかし，同時に換言を多く含んだ模倣レポートの類似度も下がってしまっている．これも，実験 1 と同様に換言への未対応が原因と考えられる．

また，文書間類似度に動的計画法を用いると，文・段落単位での入れ替えに対応出来なくなってしまうことは明らかである．これより，動的計画法の導入は全編を通した，丸写しの模倣レポート判別でならば有効である手法と考えられる．

8 おわりに

本研究では模倣レポートの判別に用いる文書間類似度を提案し，その手法を実装し，実験用のレポート集合に対して実験を行った．

本手法の特徴は，

- 単語の意味的な類似度と文間類似度を切り離し，単語の「関係」で類似度を求めることで，たとえ同一分野について書かれている文でも，文間類似度を確実に抑制している．
- 単語の対応に換言などの従来研究を柔軟に取り込むことができる

そして，実験用のレポート集合に対して実験を行った結果，換言への対応はしていなくとも，模倣したレポートを判別できる結果となった．よって，本手法で模倣レポートを高い精度で分類できると考えられる．

今後，換言への対応を行い，文献 [1], [2], [3] の手法や，文献 [5], [6] にある他の文書間類似度との比較を行っていく．また，本手法の実験・考察を行うには，データが少なすぎるため，著作権などの問題をクリアした,「生の」データを収集する必要がある．

謝辞

本研究は文部科学省 21 世紀 COE プログラム「インテリジェント ヒューマンセンシング」及び，日本学術振興会 科学研究費基盤研究 (C) (2) 13680444 の援助により行われた．

また，本研究のために模倣レポートを作成してくれた元豊田工業高等専門学校の友人たちに感謝する．

参考文献

[1] 村田 哲也，黒岩 丈介，高橋 勇，白井 治彦，小高 知宏，小倉 和久，学生レポートの n-gram による類似度評価の検討，情報科学技術フォーラム（FIT）2002 講演論文集，pp.101-102 (2002).

[2] 小川 貴博，岩堀 祐之，岩田 彰，情報メディア教育における類似レポート判定システムの構築，平成 13 年度電気関係学会東海支部連合大会講演論文集，604，p.304 (2001).

[3] 深谷 亮，山村 毅，工藤 博章，松本 哲也，竹内 義則，大西 昇，単語の頻度統計を用いた文章の類似性の定量化，電子情報通信学会論文誌，Vol.J87-DII No.2，pp.661-672 (2004).

[4] 金 明哲，助詞の分布に基づいた日記の書き手の識別，計量国語学 Vol.20 No.8，pp.357-367 (1997).

[5] 高橋 哲郎，乾 健太郎，松本 裕治，テキストの構文的類似度の評価方法について，情報処理学会 自然言語処理研究会 NL-150-24，pp.163-170 (2002).

[6] 鈴木 潤，平尾 努，佐々木 裕，前田 英作，階層構造を利用したテキスト間類似度の効率的計算法，情報処理学会 自然言語処理 NL-154，pp.101-108 (2003).